In: A. Zampolli, A. Capelli (eds., 1984): The possibilities and limits of the computer in producing and publishing dictionaries. Linguistica Computationale III, Pisa: Giardini, 279-288

MULTIFUNCTIONAL DICTIONARIES

HARALD H. ZIMMERMANN

The cost of developing a monolingual or multilingual dictionary is extremely high. On the other hand, pieces of lexical information are equally needed for various purposes. With the application of Electronic Data Processing and in particular with the setting up of large-scale data bank systems we are, technologically speaking, well equipped to produce a multifunctional lexicon system which is capable of handling the most diverse application possibilities both in theory and in practice. In the article the following points, amongst others, are dealt with:

1. Base system of a multifunctional dictionary
2. System of dictionary entries (thesaurus components)
3. Data protection and questions of data security
4. Interfaces (utilization, products)
5. Consideration of user-specific data
6. Uses of a multifunctional dictionary system

1. Base system of a multifunctional dictionary

A computerised dictionary is interpreted in the following analogue to the general conception of information processing - as a data bank. Because the stored information is to be frequently interlinked, it is not sufficient here to merely employ a hierarchically orientated system. It is, actually, rather more based on a network model. This is regarded as being equivalent to a relational data bank.

In the basic data bank all the information which is stored in the dictionary is both technically and logically speaking available. In this way it should, above all, be possible to produce data which is both free from contradictions and which is consistent.

On the other hand, as we know, natural entities cannot be - often as a result of varying cultural constraints - directly related to one another. The base system must be capable

of describing and producing even varying or more or less vague concepts. This is however more a linguistic-logic than a software technology problem.

2. System of dictionary entries

A traditional dictionary is usually recognised as having certain specially marked symbol sequences as key words which are then used to access further information. In doing this, one key is normally regarded as being a «representative» of a set of words or word forms with their information («lemma name»), and it is expected that the user will carry out these steps before he begins to look up a word in a dictionary. It is however possible to think of or even physically «see» dictionaries of a different kind (e.g. word form dictionaries, lemmatized dictionaries, concordances) in which this classification only takes place within the dictionary itself.

Another general principle of traditional and computerized dictionaries is that of pointers. A pointer element can here again be interpreted as a key word, by which it is then possible to access further information. These pointers may take the form of specific indices. If it is a question of semantic relationships between proposed entries, then *thesauri* or appropriate word families are created.

3. Data protection and questions of data security

A data bank system offers, as a rule, the best method of protection for data or segments of information against *unauthorised access.* At the same time, the *technical* safety facilities against accidental destruction are more reliable. Above all, what must be emphasised here is the function of protection against unauthorised access. Without a complex protection and security system it is hardly conceivable that firms and concerns of various kinds will be in a position to use, on the one hand, (money) saving possibilities - which are possible by using data from foreign sources - or, on the other hand, to use their own data, which otherwise can only be used and employed by other organisations paying the necessary *royalties.*

4. Interface to usage and application

In a similar way to which the «view» on *a system* is distinguished from the «view» of the *individual users of* the respective data, so should the system of a multifunctional dictionary, from the user's point of view, become somewhat diverse. In addition to this the *base system* must provide corresponding interfaces which then occur along with various possibilities of an on-line use of either the whole system or subsystem - or program systems which then put the respective (sub)user in a position to achieve his specific aims, helped by the data from the base system.

In doing so the user can then apply the system indirectly - (e.g. obtain access to his selected entries and bits of information for a computer-aided text analysis) - but in addition there should also exist the possibility of obtaining *public orientated* dictionaries (monolingual or multilingual) or of producing specialised dictionaries for use in documentation.

5*. Consideration of user specific data*

The «view» on the data, but also the requirements of the user can render it necessary that information be brought in which is *only* relevant from this particular angle. A «public orientated» definition or explanation of a term may, for example, be quite different from a definition from a special branch of science.

Leading on from this, it must then be possible to allow for these requirements by applying a corresponding labelling system to the entries or bits of information. In the extreme case for example, where there is complete divergence of the entry content dependent on the user and language spectrum - we come across physically separated «dictionaries». But this must surely be the exception to the rule rather than a common occurrence.

6. Uses of a multifunctional dictionary system

If one takes all the specifications into account which can arise as a result of varying user methods (automated language processing, terminology dictionaries, thesauri, standard language and specialised language dictionaries, text dictionaries) - one sees that the advantages of such a cumulative integrating system outweigh the disadvantages: vast reduction in costs due to the compiling and opening up of lexical information (avoidance of unnecessary work), saving on program and computing time (by providing corresponding service functions), but also the expanding network of scientific and production orientated ventures, made possible by the integration of the system. This must not be allowed to *hinder* individual ventures or research.

The system must therefore provide sufficient scope for the user to reap the benefits of the system without limiting him in his work. Using modern storage and databank technology such a concept is quite conceivable.

In a similar way to which related networks of the most varying kinds exist today for specialised information and documentation, so could the exchange of information in a lexical information system be organised.

In the following, some examples of applications are given.

6.1. Machine Lexicography for Public Use

This area of application is in the broadest sense based upon the traditional processes and structures of dictionary production. Possible users include schoolchildren, trainees, journalists and so-called «interested laymen». The information to be included ranges from spelling, stress, pronunciation of a word, through inflexion, details of etymology and semantic distinctions to details of word class and other syntax. A good example of this is the *Große Deutsche Worterbuch* by Wahrig/Brockhaus. Here more than 220,000 keywords with more than 550,000 definitions are collected and structured in such a way that they can be read and processed by computer. In a comparable way to this *monolingual* dictionary, general language dictionaries, which are intended to be bilingual, can also be developed.

With the integration of an application of this kind, interfaces, above all, are to be created for the selection of *paper* products by photo-composition. Specific dictionaries (about spelling, pronunciation, etymology and so forth) can easily be generated automatically as required from this set of base data through the inclusion/omission of certain details. With suitable maintenance (updating) and marking, dictionaries about style and neologisms can be compiled.

In the medium term (during the second half of the 1980s) more and more homes and offices will have a terminal with which it will be possible to access this general information on-line. Corresponding opportunities already exist at present e.g. through videotex. Here the public telephone network is used, in order to be able to gain access to specialised data banks. The video disc will present a definite alternative to this, on which dictionaries of this kind (perhaps more concentrated or expanded to include still/moving pictures) can likewise be offered.

6.2. Text and Language Processing in the Office

Text processing in the office will especially profit from a series of additional information. In the 1980s these will at first be very simple but nonetheless very effective aids to above all rationalisation.

The *automatic separation of syllables* (hyphenation) and the automatic split-line connected with it, above all, with languages with word composition (German, Dutch, Danish, Finnish,...) will thus profit from lexicalisation. In the case of the conversion of information stored in the base dictionary concerned (exceptions in the separation of syllables) and of word decomposition through use of base form entries as well as details of inflexion, the problem of automatic hyphenation can be regarded as solved for rudimentary cases (word ambiguity with varying separation).

It is similarly the case with *automatic spelling mistake recognition and,* should the case arise, *correction.* On the basis of information (even about word compounding with morphemes) available in the base lexicon, much more reliable processes for spelling

mistake recognition can be developed than is possible with for example the help of statistical processes (transition probability of graphemes). In German the details about word stems in the base lexicon can be utilized for the control of upper and lower case printing. A system of rules can be developed which enables one to suggest corrections for words spelt incorrectly - as far as no substitutes arise - or even to correct the mistake automatically.

A huge improvement in text systems of this kind can be prognosticized through the realisation of these two «intelligent» functions in text processing in offices, administrations, and publishing houses on the basis of information stored in a base lexicon (admittedly representative for one language) alone.

6.3. Computer-aided Language Translation

In computer-aided language translation the following areas can be roughly defined:

1) Intellectual translation through the use of a computer dictionary.
2) Semi or fully automatic translation using a system which uses rules for translation as well as a computer dictionary.

6.3.1. Use of (Multilingual) Computer Dictionaries

Once computer dictionaries (in this case the base lexicon) could only be used in the compilation and updating of (specialised) terminology, the output (on a microfiche or paper version generated from the base data) was used by the (human) translator. For this purpose, the coordination of lexical equivalents and subject area markings were sufficient in the base lexicon; in some cases semantic (thesaurus-like) networks were also useful.

In the case of an *on-line use* the utilization of the morphosyntactic functions of the base lexicon would also be required, above all when potential translation equivalents are to be extracted automatically from computer-stored data. A special realisation will be the

computer pocket dictionary (above all the travel pocket dictionary). They will profit at first from the word inventory, conversely though the information in the base system will be «enriched» with details of the area to which a term is specific (e.g. «food and drink», «train journey», «booking of a room»). These systems, which today still are insufficient, will (at first in a limited way) profit from the grammar information in the base systems, in that they use these in the generation of «short sentences» from idioms through the use of variables.

6.3.2. Rule Orientated Language Translation

A larger-scale application of the computer in language translation will have to serve almost the whole inventory of the base lexicon, from the morphological through syntactic to semantic information. Word relationships (semantic networks) and subject area markings for the disambiguation of word ambiguities (and consequently for the improvement of the translation) will be used herewith. This question will have a considerable effect on the differentiation of dictionary entries and will certainly affect the applications.

6.4. Machine Documentation, Text Analysis, and Retrieval

Although a large part of the functions discussed in (6.3.2) are valid in the area of automatic text analysis and the retrieval of texts or documents, this function is executed here separately. The reason for this of late has been that text analysis is designed for representation and condensing (i.e. avoidance of redundancy) and less for the 1:1 conversion of linguistically coded knowledge.

The *intellectual* indexing and text classification will already profit from the data stored in a base lexicon. In connection with this it is worth mentioning the semantic relationship function. It would be sufficient to in addition mark the membership to an indexing system and, if the case should arise, to introduce the remaining relation pair «used for/use for». In a similar way classification systems (if necessary, even differing ones) can be integrated. This can happen either directly through information about a classification or indirectly through a hierarchical network of relevant terms.

Machine or (semi) automatic indexing will again profit from nearly *all* information which is contained in a multifunctional base lexicon. According to past experience the morpho-syntactic information (including information about the useful decomposition of a compound and derivational relationships) and the identification of idioms/concepts with several words are especially important. Through the thereby possible syntactosemantic *lemmatisation* (which is executed by a machine language analysis system) the at present still standard word form oriented textual information retrieval systems will be superceded in the near future. Functions like truncation (i.e. the omission of the remainder of a word or a series of characters in the identification of a text descriptor), but also the ominous «WITH» and «SAME» functions will then be able to be replaced by *more useful* ones such as «WORD PART» or «ATTRIBUTION», or even replaced by *relevance functions of* the system which further «unburden» the user of rather «technical» details in his enquiry (cf. e.g. the CONDOR and CTX systems which have already been tested in prototype form).

Even «more complex» or «more far-reaching» systems for handling language (for example the so-called Question-Answering Systems in the area of artificial intelligence research) will profit from the information contained in the multifunctional lexica discussed above.

As the limit to encyclopedic information (generally speaking: the transition from linguistic to technical subjectmatter knowledge) - if one wants to make any distinction at all - is looked upon as fluid, one can imagine a series of extensions of multifunctional (probably also then encyclopedic) lexica, which may have an effect upon areas of application mentioned above.

6.5. Combination with the Textual Data

In the last instance it will be interesting to link the lexical data directly with (textual) «collections of knowledge». In this case the base lexicon contains not only information *abstracted* from world knowledge and condensed by lexicographers, but forms the

interface (access key) to the primary information itself. Finally two examples to this end will be given:

1. The human translator does not only provide (possible) translation equivalents for a word, but perhaps also the contexts (or some of them) in which a term was used with this or that translation. His ultimate decision is also stored and incorporated into the network and can be used for further terminology work or text translation and so forth.
2. The results of the consideration of the standard language on the basis of so-called text corpuses but also suitable data about poetry, specific prose and so forth are fully integrated «lexically» into the base system network. Thus it will be possible for example to generate frequency dictionaries or to obtain references on-line about the use of the word «Frühling» by Goethe or «Solidarität» in the newspaper «Die Welt» (with details of the source and an arbitrary context).

One can well imagine the other kinds of possibilities which arise from this.

7. The Multifunctional Dictionary - a Utopia?

From the view of technical possibilities a multifunctional dictionary as outlined here, can certainly be realised. The stability which is perhaps still lacking today here and there in the technology (insufficient provision of characters in data base systems, speed of access, storage compactness) will judging from past experience soon be overcome. Social and psychological barriers will be able to be drastically reduced in the medium term on the basis of the *large economic* advantages when the question of data protection and data security has been satisfactorily answered. One can envisage two methods of putting this into practice:

1. «Centralised» availability of the lexicon information with the necessary software to manufacture the varied products, centralised service (e.g. via an on-line link).This variant will in all cases play a definite role in the construction phase, as long as the functions and technology necessary for variant (2) are not available. Against

this centralisation there is the danger of the monopolisation of information.

2. «Decentralised» availability of the lexicon information. This can on the one hand mean that the participants in the «lexical information alliance» are linked with one another in a net-like fashion, so that the data (which exists physically stored only *once)* is «shared» (using a so-called shared data base system). It can also mean that *physical copies* can be produced from the *original* data or parts therefrom (downloading).

The organisation of a multifunctional dictionary system of this kind - be it regional, national or international - will raise yet more questions, e.g. legal questions («Who owns which information?») or even cost problems. In view of the many problems still unsolved it seems to make little sense to develop a system of this kind artificially. This would indeed lead to a situation where it would in the final analysis be a utopia.

There is a real change however, if it is attempted *to incorporate already existing or collected lexical data* (in machine readable *form) cumulatively* into such a system. *In addition to* the system development, updating and maintenance work would indeed in this case have to be incorporated, which would make it possible to standardize the data stocks which are certainly heterogenous in small areas (as long as there are no underlying differences in prototypes and concepts, the protection of which is also the job of the system).

This realisation should as far as possible - no matter how many problems there are - take place on an international level, so that public bodies as well as the (information) industry could participate.

In order to introduce the necessary means into the investment phase - in the medium term the system seems totally able to be commercialized - products of this kind which allow an economisation of language research itself could be intensely introduced in the construction phase. The national as well as the international sponsoring institutions could be asked about the joint financing of this.

With a stepwise construction centred on small areas it can be prevented that this envisaged concept remains a utopia.